FEDERICO BUSATO

🌐 Italian   📍 Santa Clara, CA, US

🐦 @fedebusato   💼 federico-busato

PROGRAMMING LANGUAGES

`C++` and its modern versions (C++11/14/17/20/23)

`CUDA` and low-level programming PTX, SASS

WORK AND RESEARCH INTERESTS

Sparse Linear Algebra · Graph Analytics · GPU Architectures · Parallel Computing · Software Engineering · Performance Optimization

WORK EXPERIENCE

⬢ Senior Software Engineer, IC5
NVIDIA Corporation, CUDA HPC Math Libraries Team

📅 May 2018 – Ongoing   🕐 6 Years   📍 Santa Clara, CA, US

Lead Engineer of the NVIDIA sparse linear algebra group. Projects and contributions:

- `cuSPARSE` is the most popular and widely used sparse linear algebra library for GPU architectures that allows obtaining peak performance for matrix operations involving a high ratio of structural non-zero elements. The library supports several software and hardware platforms, including Linux, Windows, QNX, L4T, x86-64, IBM PowerPC, and Arm64. cuSPARSE is part of the NVIDIA CUDA Toolkit which is downloaded 8 million times per year.
  → *Contribution*: Sole engineer for four years, rewrote most of the library from pure C to C++14, worked on any aspect of the development

- `cuSPARSELt` is a specialized library to fully exploit advanced NVIDIA GPU features, such as Sparse Tensor Cores, and accelerate matrix-matrix operations in which at least one operand is a sparse matrix.
  → *Contribution*: Created the library from scratch

- `Top500 HPCG Benchmark` Together with the *High Performance LINPACK (HPL)* and the *High Performance Linpack (HPL)* benchmarks, the *High Performance Preconditioned Conjugate Gradient (HPCG)* is a fundamental tool for evaluating and ranking computer system performance. The benchmark solves a large sparse linear system using an iterative method. Contrary to the other two benchmarks, HPCG is designed to better represent many important scientific calculations, with low computation-to-data-access ratio.
  → *Contribution*: supervision

- `NVPL Sparse` Sparse linear algebra library for Arm64 Grace platform
  → *Contribution*: Interface proposal, supervision

`cuSPARSE` and `cuSPARSELt` libraries have significant importance in classical HPC applications and deep learning context. The users span from academics to the world's top tech companies and national research laboratories.

Besides algorithm design and software development, my actives include:

- Set libraries goals, roadmap, and interfaces
- Direct interaction with important customers such as Google, Amazon, Meta, Baidu, Mathworks, etc.
- Involved in the definition of future hardware features for sparse computation
- Coordination with other teams
- Involved in the hiring process. Did 100+ interviews
- Mentor new hires and interns. Mentored 7 engineers
- Provide user support through official NVIDIA developer forum, GitHub, etc.
- Write public documentation and examples

🏛 University Teaching Professional, Master's Degree course *"Modern C++ Programming"*
University of Verona, Dept. of Computer Science

📅 2018 – 2022   🕐 5 years   📍 Verona, Italy / Remote

The course covers fundamental to advanced concepts of the language and its modern extensions (C++11/14/17/20/23). It also deepens non-functional aspects of C++ projects such as improving the quality, readability, the best practices, as well as the ecosystem and specific tools. Besides University teaching, it is adopted as an important resource for interns/new hires training at NVIDIA. The course comprises over 1,500 slides, and I keep it updated year after year.

🏛 University Teaching Professional, Bachelor's Degree course *"Operating System"*
University of Verona, Dept. of Computer Science

📅 2018   🕐 1 year   📍 Verona, Italy

The course aims at providing an introduction to the principles and design of operating systems, with an emphasis on the synchronization of processes and the management of resources. The course topics are assisted by Bash and C programming to show the concepts in practice.

Software Engineer Intern
NVIDIA Corporation

📅 Oct 2017 - Jan 2018   🕐 4 months   📍 Santa Clara, CA, US

The internship focuses on the design and the implementation of GPU-accelerated time series forecasting techniques, with emphasis on AutoRegressive Integrated Moving Average (ARIMA) model. The second project involved sparse linear algebra with the development of two new strategies for sparse matrix transposition. The code has been then integrated in the public cuSPARSE library

---

🏛 Teaching Assistant, Master's Degree course *"Advanced Architectures"*
University of Verona, Dept. of Computer Science

📅 2015 - 2017   🕐 3 years   📍 Verona, Italy

The course aims at providing theoretical and practical knowledge about programming and analysis of advanced computational architectures, with emphasis on multiprocessor and GPU platforms. The course covers the main parallel computing technologies: OpenMP, MPI, CUDA, OpenCL, OpenACC.

EDUCATION

🏛 Postdoctoral Research Associate
University of Verona, Dept. of Computer Science

📅 2018   📍 Verona, Italy

🏛 Visiting Scholarship
Georgia Institute of Technology

📅 Jan 2017 - May 2017   📍 Atlanta, GE, US

Project Title: *"Streaming Graph Processing on GPU and Multi-core Architectures"*.
Advisors: Prof. David A. Bader, Oded Green
Description: The project focused on the design and the implementation of a high-performance graph framework for GPU architectures that allows the graph structure to dynamically evolve over time in an efficient way. The outcome of the visiting scholarship is the article *"Hornet: An Efficient Data Structure for Dynamic Sparse Graphs and Matrices"*, published in the IEEE High-Performance Extreme Computing Conference (HPEC). The framework is still maintained and used for research purposes with several works relying on it.

🎓 Ph.D. in Computer Science
University of Verona, Dept. of Computer Science

📅 2015 – 2017   📍 Verona, Italy

Thesis title: *"High-Performance and Power-Aware Graph Processing on GPUs"*
Advisor: Prof. Nicola Bombieri

🎓 Batchelor and Master Degree in Computer Science and Engineering
University of Verona, Dept. of Computer Science

📅 2007 – 2014   📍 Verona, Italy

Thesis Title: *"BFS-4K: an Efficient Implementation of BFS for Kepler GPU Architectures"*
Final grade: Full marks (110/110) and honour (Average 29.1/30, USA Grade A+, GPA 4.00)
Advisor: Nicola Bombieri

PROJECTS

- **NVIDIA NVPL Sparse**: High-performance Sparse Linear Algebra Library for Grace Arm64
  `docs.nvidia.com/nvpl/_static/sparse/index.html`
- **NVIDIA cuSPARSELt**: High-performance general matrix-matrix operations leveraging Sparse Tensor Core
  `docs.nvidia.com/cuda/cusparselt/index.html`
- **NVIDIA cuSPARSE**: GPU-Accelerated sparse linear algebra bibrary
  `docs.nvidia.com/cuda/cusparse/index.html`
- **NVDIA CUDA Toolkit Samples**: CUDA Library Samples contains examples demonstrating the main features and APIs of the CUDA Math Libraries ★ **1.1K+ Stars**
  `github.com/NVIDIA/CUDALibrarySamples`
- **Modern C++ Programming**: Medium to advanced C++ Course, 1500+ pages ★ **9.8K+ Stars**
  `github.com/federico-busato/Modern-CPP-Programming`
- **Hornet/HornetsNest: High-Performance Streaming Graph Analytics on GPUs**, Graph data structure and algorithms for sparse dynamic graphs and matrices, Main developer and creator, 2017
  `github.com/hornet-gt/hornet`

PUBLICATIONS

👥 Conference Proceedings

- Fraccaroli, Leonardo; Giugno, Rosalba; Cancellieri, Samuele; Busato, Federico; Bombieri, Nicola, *"FAST-CON: a Multi-source Approach for Efficient ST Connectivity on Sparse Graphs"*, In Proc. of IEEE High Performance Extreme Computing (HPEC), virtual conference, September 25-29, 2023, pp 1-6.
- Busato, Federico; Bombieri, Nicola, *"Configuring Graph Traversal Applications for GPUs: Analysis and Correlation of Implementation Strategies with Graph Characteristics"*, In Proc. of IEEE High Performance Computing Systems Conference (HPCS, APPMM), Dublin, Ireland, July 15-19, 2019, pp 1-8.
- Busato, Federico; Green, Oded; Bombieri, Nicola; Bader, A. David, *"Hornet: An Efficient Data Structure for Dynamic Sparse Graphs and Matrices on GPUs"*, In Proc. to IEEE High Performance Extreme Computing Conference (HPEC), Waltham, USA, September 25-27, 2018, pp. 1-6.
- Busato, Federico; Bombieri, Nicola, *"Efficient Load Balancing Techniques for Graph Traversal Applications on GPUs"*, In Proc. of 24th International European Conference on Parallel and Distributed Computing (EuroPar), Turin, Italy, August 27-31, 2018, pp. 1-12.
- Green, Oded; Fox, James; Kim, Euna; Busato, Federico; Bombieri, Nicola; Lakhotia, Kartik; Zhou, Shijie; Singapura, Shreyas; Zeng, Hanqing; Kannan, Rajgopal; Prasanna, Viktor; Bader, David, *"Quickly Finding a Truss in a Haystack"*, In Proc. of IEEE High Performance Extreme Computing Conference (HPEC), IEEE/Amazon/DARPA Graph Challenge,Waltham, USA, September 12-14, 2017, pp. 1-7, *Innovation Award*.
- Bonnici, Vincenzo; Busato, Federico; Akhmedov, Murodzhon; Caligola, Simone; Cascione, Luciano; Montemanni, Roberto; Fummi, Franco; Bertoni, Francesco; Bombieri, Nicola; Kwee, Ivo; Giugno, Rosalba, *"cuRnet: an R Package for the Single-source Shortest Paths Analysis on GPUs"*, In Proc. of Bioinformatics Italian Society (BITS), Cagliari, Italy, July 5-7, 2017.
- Busato, Federico; Bombieri, Nicola, *"A Performance, Power, and Energy Efficiency Analysis of Load Balancing Techniques for GPUs"*, In Proc. of IEEE International Symposium on Industrial Embedded Systems (SIES), Toulouse, France, 14-16 June, 2017, pp. 1-10.
- Bombieri, Nicola; Busato, Federico; Fummi, Franco, *"Power-aware Performance Tuning of GPU Applications Through Microbenchmarking"*, In Proc. of ACM/EDAC/IEEE Design Automation Conference (DAC), Austin, Texas, USA, June 18-22, 2017, pp. 1-6, *HiPEAC Award*
- Aldegheri, Stefano; Barnat, Jiri; Bombieri, Nicola; Busato, Federico; Ceska, Milan, *"Parametric Multi-Step Scheme for GPU-Accelerated Graph Decomposition into Strongly Connected Components"*, In Proc. of International European Conference on Parallel and DistributedComputing (Europar), Workshop on Performance Engineering for Large Scale Graph Analytics (PELGA), Grenoble, France, August 24-26, 2016, pp. 1-12.
- Bombieri, Nicola; Busato, Federico; Fummi, Franco; Scala, Michele, *"MIPP: A Microbenchmark Suite for Performance, Power, and Energy Consumption Characterization of GPU architectures"*, In Proc. of IEEE International Symposium on Industrial Embedded Systems (SIES), Krakow, Poland, May 23-25, 2016, pp. 1-8.
- Bombieri, Nicola; Busato, Federico; Fummi, Franco, *"A Fine-grained Performance Model for GPU Architectures"*, In Proc. of ACM/IEEE International Conference on Design, Automation and Test in Europe (DATE), Dresden, Germany, March 14-18, 2016, pp. 1-8.

- Bombieri, Nicola; <u>Busato, Federico</u>; Danese, Alessandro; Piccolboni, Luca; Pravadelli, Graziano, *"Exploiting GPU Architectures for Dynamic Invariant Mining"*, In Proc. of IEEE International Conference on Computer Design (ICCD), New York City, NY-USA, October 18-21, 2015, pp. 192-195.
- <u>Busato, Federico</u>; Bombieri, Nicola, *"On the Load Balancing Techniques for GPU Applications Based on Prefix-scan"*, In Proc. of IEEE International Symposium on Embedded Multicore/Manycore System-on-Chip (MCSoC), Turin, Italy, September 23-25, 2015, pp. 88-95.
- Bombieri, Nicola; <u>Busato, Federico</u>; Fummi, Franco, *"An Enhanced Profiling Framework for the Analysis and Development of Parallel Primitives for GPUs"*, In Proc. of IEEE International Symposium on Embedded Multicore/Many-core System-on-Chip (MCSoC), Turin, Italy, September 23-25, 2015, pp. 1-8.

📖 Journal Articles

- Mocci, Jacopo; <u>Busato, Federico</u>; Muradore, Riccardo; Bombieri, Nicola, *"Efficient Implementation of the Shack-Hartmann Centroid Extraction for Edge Computing"*, in Journal of the Optical Society of America (OSA), 2020, pp 1-9.
- <u>Busato, Federico</u>; Danese, Alessandro; Piccolboni, Luca; Pravadelli, Graziano; Bombieri, Nicola, *"Mangrove: an Inference-based Dynamic Invariant Mining for GPU Architectures"*, in IEEE Transactions on Computer (TCOMP), DOI 10.1109/TC.2019.2953846, 2019, pp. 1-14.
- Bonnici, Vincenzo; <u>Busato, Federico</u>; Aldegheri, Stefano; Akhmed, Murodzhon; Cascione, Luciano; Carmena, Alberto Arribas; Bertoni, Francesco; Bombieri, Nicola; Kwee, Ivo; Giugno, Rosalba; *"cuRnet: an R Package for Graph Traversing on GPU"*, In BMC Bioinformatics, 2018, pp. 1-9.
- <u>Busato, Federico</u>; Bombieri, Nicola, *"A Dynamic Approach for Workload Partitioning on GPU Architectures"*, In IEEE Transactions of Parallel and Distributed Systems (TPDS), Vol. 28, Issue 6, June 2017, pp. 1-14
- <u>Busato, Federico</u>; Bombieri, Nicola *"An Efficient Implementation of the Bellman-Ford Algorithm for Kepler GPU Architectures"*, In IEEE Transactions of Parallel and Distributed Systems (TPDS), Vol. 27, no. 8, pp. 2222-2233, 2016.
- Bonnici, Vincenzo; <u>Busato, Federico</u>; Micale, Giovanni; Giugno, Rosalba; Pulvirenti, Alfredo; Bombieri, Nicola, *"APP-AGATO: an APproximate PArallel and stochastic GrAph querying TOol for biological networks"*, In Bioinformatics, pp. 2159-2166, 2015.
- Bombieri, Nicola; <u>Busato, Federico</u>; Fummi, Franco, *"Pro++: A Profiling Framework for Primitive-based GPU Programming"*, In IEEE Transactions on Emerging Topics in Computing (TECT), pp. 1-12, 2015.
- <u>Busato, Federico</u>; Bombieri, Nicola, *"BFS-4K: an Efficient Implementation of BFS for Kepler GPU Architectures"*, In IEEE Transactions of Parallel and Distributed Systems (TPDS), Vol. 26, no. 7, pp. 1826-1838, 2015.

📓 Book Chapter

- <u>Busato, Federico</u>; Bombieri, Nicola, *"Graph Algorithms on GPUs"*, in "Advanced in GPU Research and Practice", Elsevier, in printing 2016.

AWARDS

- 🎖 **DAC HiPEAC Award**, Bombieri, Nicola; <u>Busato, Federico</u>; Fummi, Franco; *"Power-aware Performance Tuning of GPU Applications Through Microbenchmarking"*, In Proc. of ACM/EDAC/IEEE Design Automation Conference (DAC), USA, June 18-22, 2016, pp. 1-6.
- 🎖 **Innovation Award IEEE/Amazon/DARPA Graph Challenge** Green, Oded; Fox, James; Kim Euna; <u>Busato, Federico</u>; Bombieri, Nicola; Lakhotia, Kartik; Zhou, Shijie; Singapura, Shreyas; Zeng, Hanqing; Kannan, Rajgopal; Prasanna, Viktor; Bader, David, *"Quickly Finding a Truss in a Haystack"*, In Proc. of IEEE High Performance Extreme Computing Conference (HPEC), Waltham, USA, September 12-14, 2017, pp. 1-7

PATENT

- 📄 **Workload Assignment Technique**, 22-SC-1045US01 / 0112912-720US0 (pending).
  Summary: Optimal fine-grained CSR/CSC Load Balancing

PRESENTATIONS

🖥 Oral Presentations

- *"Just-In-Time Link-Time Optimization Adoption in cuSPARSE/cuFFT: Use Case Overview"*, NVIDIA GPU Technology Conference (GTC), November 8-11, 2021
- *"High-Performance Sparse Linear Algebra on NVIDIA GPUs with cuSPARSE"*, Invited, Oral Presentation at SIAM Conference on Computational Science and Engineering (CSE21), Virtual Conference, March 1-5, 2021.
- *"Configuring Graph Traversal Applications for GPUs: Analysis of Implementation Strategies and their Correlation with Graph Characteristics"*, Oral Presentation at 17th High-Performance Computing Systems Conference (HPCS, APPMM), Dublin, Ireland, July 15-19, 2019.
- *"Efficient Load Balancing Techniques for Graph Traversal Applications on GPUs"*, Oral Presentation at International European Conference on Parallel and Distributed Computing (EuroPar), Turin, Italy, August 27-31, 2018.
- *"A Performance, Power, and Energy Efficiency Analysis of Load Balancing Techniques for GPUs"*, Oral Presentation at IEEE International Symposium on Industrial Embedded Systems (SIES), Toulouse, France, June 14-16, 2017.
- *"Parametric Multi-Step Scheme for GPU-Accelerated Graph Decomposition into Strongly Connected Components"*, Oral Presentation at International European Conference on Parallel and Distributed Computing (Europar), Workshop on Performance Engineering for Large Scale Graph Analytics (PELGA), Grenoble, France, August 24-26, 2016.
- *"MIPP: A Microbenchmark Suite for Performance, Power, and Energy Consumption Characterization of GPU architectures"*, Oral presentation at IEEE International Symposium on Industrial Embedded Systems (SIES), Krakow, Poland, May 23-25, 2016.
- *"A Fine-grained Performance Model for GPU Architectures"*, Oral presentation at ACM/IEEE International Conference on Design, Automation and Test in Europe (DATE), Dresden, Germany, March 17, 2016.
- *"On the Load Balancing Techniques for GPU Applications Based on Prefix-scan"*, Oral presentation at IEEE International Symposium on Embedded Multicore/Many-core System-on-Chip (MCSoC), Turin, Italy, September 23, 2015.
- *"An Enhanced Profiling Framework for the Analysis and Development of Parallel Primitives for GPUs"*, Oral presentation at IEEE International Symposium on Embedded Multicore/Many-core System-on-Chip (MCSoC), Turin, Italy, September 23, 2015.

BLOG POSTS

Official NVIDIA Developer Blog Posts
- **Accelerating Matrix Multiplication with Block-Sparse Format and Tensor Cores**, March 19, 2021
  developer.nvidia.com/blog/accelerating-matrix-multiplication-with-block-sparse-format.
- **Exploiting NVIDIA Ampere Structured Sparsity with cuSPARSELt**, December 8, 2020
  developer.nvidia.com/blog/exploiting-ampere-structured-sparsity-with-cusparselt/

MENTORING

- **San Diego Supercomputer Center (SDSC) GPU Hackathon 2022**, May 3, May 10-12, 2022
- **Princeton GPU Hackathon 2022**, June 1, June 6-8, 2022
- *"NVIDIA Math Libraries: cuSPARSE and cuSOLVER"*, **Connect with the Experts Session at NVIDIA GPU Technology Conference (GTC)**, March 21-24, 2022
- *"Tensor Core-Accelerated Math Libraries for Dense and Sparse Linear Algebra in AI and HPC"*, **Connect with the Experts Session at NVIDIA GPU Technology Conference (GTC)**, April 12-16, 2021

EXTERNAL COLLABORATORS

- **Wen-mei Hwu**, University of Illinois, Urbana-Champaign
  Research collaboration on sparse linear algebra, 2021

LANGUAGES

Italian   English